Using BERT to do feature extraction and Drug-Drug interaction prediction

Meng Rui (122090405)
School of Data Science
Chinese University of Hong Kong, Shenzhen
ruimeng@cuhk.edu.cn

Abstract

The main topic for this report is about "using speech and language processing techniques to solve a real-world problem". So I came up with the idea of using BERT to realize a drug-drug interaction prediction task. Bert (Bidirectional Encoder Representation from transformers) is a pre-trained deep learning model designed for natural language understanding tasks. In medical research, natural language processing techniques can help to extract more consice information from huge text. While molecular can be represented as a series of characters, similar to natural language, they are inherently different in structure and context. Thus a pretrained BERT model was used to do feature extraction towards drug sequence. And after feature extraction we input the feature vector to a randomforest predictor to test the performance.

1 Introduction

In clinical treatments, it's common to use two or more drugs at the same time. Thus we have to test whether interaction will occur between the two drugs. While wet lab experiments is time-costing and costly, computational method have emerged as a key tool for large DDI scale prediction. So far, the computational approach can mainly divided into three catogories: score-function based method, machine learning method and deep-learning method. Score function based method mostly do prediction work based on algorithmic principles, and this kind of method doesn't need negative samples. Machine learning method are suitable for few prior knowledge, as the model can learn features from the provided data. And deep learning methods also become a popular method for DDI prediction. As with the combination of NLP and pattern recognition, models can learn from input features and allow the extraction of multiple features from drug pairs. In this assignment, I create a prediction model based on machine learning based method. The feature of drug pairs was extracted from SMILES and knowledge graph from , then we do the prediction work with RandomForest model.

2 Related Work

Huang et al. introduced a novel metric, "S-score," designed to evaluate the strength of network connections between drug targets, aiming to determine potential drug interactions [1].
Takeda et al. explored multiple dimensions of drug similarity and developed a logistic regression model based on calculated similarity scores to predict binary drug-drug interactions (DDIs) [2].
Ren et al. highlighted that incorporating multi-scale information can significantly improve the prediction accuracy of DDIs for novel drugs. They proposed a model called BioDKG-DDI, which leverages attention mechanisms to integrate diverse biochemical features and perform binary predictions on balanced datasets [3].
Su et al. concentrated on utilizing biomedical knowledge graphs (KGs) to provide comprehensive information about drug properties and associated facts. They constructed an attention-based neural network to predict binary DDIs using the data from these KGs [4].

3 Approach

For this assignment, I use a multi-scale feature extraction method to extract the molecular's feature and make prediction based on machine learning models. For feature extraction, embedding vectors of drug SMILES were generated using a pretrained graph neural network. The result feature vectors then passed to the BERT model, and finally we can get a feature vector representing drug-drug pairs. Then the pairs were fed to the RandomForest prediction model. The feature vector not only contain 2-D information from SMILES sequence, but also contain 3-D information from GNN graph, thus it's a multi-scale feature extracting model.

4 Experiments

4.1 Data

A primarily dataset from Therapeutics Data Commons (TDC) was imported for this research. Specifically, I choose DDI dataset through TDC library. The dataset includes two drug molecular representations(In SMILES), as well as a binary prediction of interaction. While the dataset only includes positive dataset, I generate negative dataset by randomly chosing one molecular, and pick out another drug molecular which shares minimum similarity with this drug. The ratio of positive and negative dataset is 1:1.The dataset used for this research was attached in data package.

4.2 Evaluation method

The evaluation metrics used for this research is AUPR, accuracy, precision score, recall and F1 score. All to evaluate the prediction model performance.

4.3 Results

The results are the evaluation outcome of the model, which are attached below:

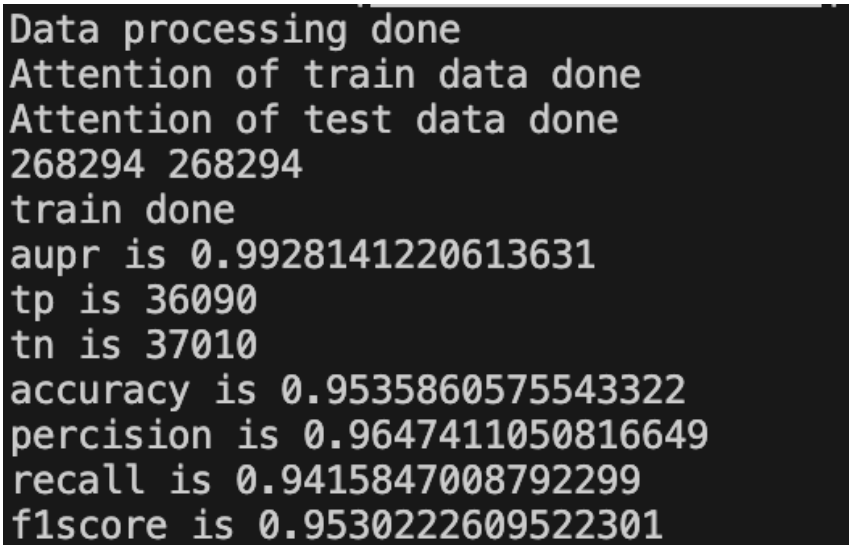

Figure 1: Evaluation results of the model.

5 Analysis

From the evaluation results, we can conclude it's a state-of-art method. The model has showed satisfying output. The AUPR score is close to 1. And the model has showed high accuracy and percision scores.

6 Conclusion

The model has showed satisfying output. I think it's mainly the reason of implementing Bert to do feature extraction and the combination of text and structural information. However, the experiments are lack of baseline comparing models. And the prediction model used for this assignment is quite simple. Also, there is no case study for this model. So further improvements needs to be done.

References

[1] X. Huang et al. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS computational biology*, 9:e1002998, 2013.

[2] Y. Takeda et al. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *Journal of cheminformatics*, 9:1–9, 2017.

[3] Z. Ren et al. Biodkg–ddi: predicting drug–drug interactions based on drug knowledge graph fusing biochemical information. *Briefings in Functional Genomics*, 21:216–229, 2022.

[4] L. Su et al. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Briefings in bioinformatics*, 23:bbac140, 2022.